Project Proposal- London Cycle Hire Data

What is the problem you want to solve?

The pubic bike hire scheme is a great way of maximising the efficiency of public transportation systems while reducing pollution. In London, the Sanatander cycle hire scheme was started in 2010 (originally sponsored by Barclays) and it continues to be an envionmentaly friendly and healthy way to explore the city or just commute. Another useful aspect of the scheme is that bikes can be hired at any time of the day and there are 839 stations densely populated around central London.

In recent years, more players are entering the cycle hire scheme such as Mobike and Ofo that offer lower pricing for short trips under 30 minutes with a more seamless experience that doesn't require docking. While we cannot change the way the bikes are tracked, we can improve overall output of the system by matching availability with demand and identifying. In a recent survey done about the public cycle hire scheme, the key consumer concerns were bike availability and space availability at docking stations (see quarterly reports released by the scheme

The key problems to solve will be:

1. Understanding the current utilization of the public hire scheme using historical data
2. Predicting the future demand for bikes at each station
3. Use the predictive model to improve availability of bikes to maximize the output(sales - costs)

Who is your client and why do they care about this problem?

In this case, my client will be Transport For London (TFL) and

What data are you using? How will you acquire the data?

TFL provides data on every single ride from 2012 up till August 2019 on an AWS file storage system. This is updated regularly through csv files reflecting data over 2 weeks. The data consists of the following features:

- Rental ID
- Bike ID
- Start time, End time, Duration
- Start Station Name, End Station Name, Start Station ID, End Station ID

There is also a seperate file that gives the location data of each station (latitude, longitude) and the number of docks at each station.

The data will be downloaded through AWS file storage system, S3 using the boto module. See AWS documentation for more information. I've included an example of downloading the data for a 2 week period in January 2016 below along with some data manipulation to get a better idea of what can be done with it.

Brief outline of how you'll solve the problem

The problem will first require the analysis of the data. The data wrangling and analysis will mostly be done through the pandas library along with various visualization libraries to enable easier understanding of the relationships between the features. It would also require adding external parameters such as weather conditions, temperature and type of day (weekday/holiday) to better understand the trends that drive cycle hire demand.

The second part of solving the problem would require selecting the appropriate features and using a supervised machine learning model to predict the demand. The data will be split into training and testing data with the aim of minimizing the test error.

What are the deliverables

The deliverables are

1. Code which produces a predictive model
2. Report detailing the methods used to solve the problem
3. Presentation that summarizes the methods used